Attorney Docket No.: 020910000410

PATENT APPLICATION

METHOD FOR SELF-VALIDATION OF MOLECULAR MODELING

Inventor:

Michael G. Hollars, a citizen of the United States of America, residing at:
1514 South Mary Avenue
Sunnyvale, California 94087; and

Michael A. Sherman, a citizen of the United States of America, residing at:
561 Bush Street
Mountain View, California 94041

Assignee:

Protein Mechanics, Inc.
278 Hope Street, Suite C
Mountain View, California 94041

Entity:

Small

METHOD FOR SELF-VALIDATION OF MOLECULAR MODELING

CROSS-REFERENCES TO RELATED APPLICATIONS

This application is entitled to the benefit of the priority filing date of Provisional Patent Application No. 60/245,734, filed 2000 Nov. 2, and in addition, co-pending Provisional Patent Application No. 60/245,730, filed 2000 Nov. 2; No. 60/245,731, filed 2000 Nov. 2; and No. 60/245,688, filed 2000 Nov. 2; all of which are hereby incorporated by reference.

BACKGROUND OF THE INVENTION

The present invention is related to the field of molecular modeling and, more particularly, to computer-implemented methods for the dynamic modeling and static analysis of large molecules.

The motion of bodies is determined by Newton's Laws of Motion. For a body subject to a force, Newton's Second Law states:

$$F = ma$$

or the acceleration a of the body is equal to the total force upon the body. This simple equation hides enormous complexity for the dynamic modeling and static analysis of large molecules. The acceleration of the body is the time derivative of velocity of the body and to determine the velocity of the body, its acceleration must be integrated with respect to time. Likewise, the velocity of a body is the time derivative of position of the body and to determine the position of the body, its velocity must be integrated with respect to time. Thus with knowledge of the force upon a body, integration operations must be performed to determine the velocity and position of the body at a given time.

In a molecule, there are multiple bodies whose motions must be considered. Each body, an atom, of the molecule is subject to multiple and complex forces. Thus the calculation of the motion and the shape of the molecule requires the determination of the position and motion of each atom of the molecule. Hence the calculation of the structure, dynamics and thermodynamics of molecules, including complex molecules having thousands of atoms, by computers would seem to be the perfect answer.

Indeed, the field of molecular modeling has successfully simulated the motion (molecular dynamics or MD) and the rest states (static analysis) of many complex molecular systems by computers. Typical molecular modeling applications have included enzyme-

ligand docking, molecular diffusion, reaction pathways, phase transitions, and protein folding studies. Researchers in the biological sciences and the pharmaceutical, polymer, and chemical industries are beginning to use these techniques to understand the nature of chemical processes in complex molecules and to design new drugs and materials accordingly. Naturally, the acceptance of these tools is based on several factors, including the accuracy of the results in representing reality, the size of the molecular system that can be modeled, and the speed by which the solutions are obtained. The accuracy of the solutions is generally accepted.

However, the validity of the models should be tested and the models refined if any modeling approximations are inappropriate. In current practice, computed results are compared with empirical results measured in the laboratory. The present invention can exploit internal consistency requirements to obtain a degree of validation from the computational method directly, without recourse to the laboratory.

SUMMARY OF THE INVENTION

The present invention provides for a method for validating a computer modeling of a molecular system. The method has the steps of selecting a model parameter of the molecular system; selecting a validation measure of the molecular system; simulating the molecular system by the computer modeling with the selected model parameter; then determining a value of the validation measure of said molecular system from the simulating step; and testing whether the value of the validation measure is in a predetermined range to validate the computer modeling. The method can be performed iteratively by varying the model parameter continuously, such as varying a temperature model parameter, or discretely, such as substituting for different residues in a protein.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flow chart of self-validation method for a molecular model, according to the present invention;

Fig. 2 is a representation of an exemplary rigid multibody system model of an alanine dipeptide in accordance with the present invention;

Fig. 3A is a detail of the Fig. 1 model to illustrate the joint reactions exerted on the peptide bond of the alanine dipeptide; Fig. 3B illustrates the addition of a pin joint to refine the peptide bond model of the alanine dipeptide in accordance with the present

invention; Fig. 3C illustrates the addition of a slider joint to refine the peptide bond model of the alanine dipeptide in accordance with the present invention;

Fig. 4 is a plot of the torque magnitude versus simulation time for a polypeptide rigid multibody system model;

Fig. 5 is a plot of the calculated minimum potential energy versus starting angle ψ of an alanine dipeptide rigid multibody model; and

Fig. 6 is a graph of the calculated minimum potential energy of a dipeptide alanine-R, where R are various discrete peptide residues interchanged.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

The computer molecular modeling may be self-validated in accordance with the present invention. Molecule modeling and simulations are made with certain approximations, such as rigid body approximations of clusters of atoms, and the parameters of the models of the force fields, solvents, initial conditions, and other environmental and internal models. Though the present invention is not necessarily limited to such molecular modeling and simulations as described in co-pending U.S. Patent Application No. _____, entitled "METHOD FOR LARGE TIMESTEPS IN MOLECULAR MODELING" and claiming priority to the above-referenced Provisional Patent Application No. 60/245,688; U.S. Patent Application No._____, entitled "Method for Residual Form in Moecular Modeling "and claiming priority to the above-referenced Provisional Patent Application No. 60/245,731; and U.S. Patent Application No. _____, entitled " "and claiming priority to the above-referenced Provisional Patent Application No. 60/245,730; all of which patent applications filed of even date, assigned to the present assignee and incorporated by reference in their entirety, the resulting high-speed molecular modeling taught therein are particularly useful in exploiting the advantages of the present invention.

Fig. 1 illustrates a flow chart of the general steps of the molecular modeling self-validation method of the present invention. In initial step 100 a molecular dynamics (MD) simulation is created with a model parameter P and a validation measure M. Parameter P is chosen to test a particular modeling assumption or approximation, such as the rigid-body modeling assumption discussed in the above referenced co-pending applications, a constant of an atomic force field or solvent model, or even structure of the model itself (the particular amino acid sequence). Measure *M* is a result of the MD simulation and is chosen to validate the modeling assumption or approximation.

The modeling parameter P is set to its initial value in step 102. The MD simulation is run in step 104 for the current setting of P, and the measure M is computed. Step 106 tests whether all the settings of P have been run. If not, then P is set to a new value by step 108, and the simulation is re-run. If all settings of P have been tested, then the testing of the validation measures occurs in step 110. With this method, different types of parameters P can be tested. The particular parameter P determines how many settings of P are required for the validation method, how the measures M are derived to test P, and exactly how the validation tests are conducted.

Two general types of validation tests can be used. In the first type, the molecular model is run with one or more settings or substitutions of the modeling parameter P: $P_1, P_2 \ldots P_i \ldots P_n$. Then the validity measure M is tested to determine whether it lies within a specified range:

$$M_{\min} < M < M_{\max}$$

If M is outside the valid range, then the model should be modified, and the validation test rerun until M falls within the desired range.

In the second type of validation test, the simulation test is run with two settings of P, i.e., P_1 and P_2, with two resulting measures of M, M_1 and M_2. Then the partial derivative of M with respect to P is tested to determine the partial derivative lies within a specified range:

$$\frac{\partial M}{\partial P} \approx \frac{\Delta M}{\Delta P} = \frac{M_2 - M_1}{P_2 - P_1}$$

$$\left(\frac{\partial M}{\partial P}\right)_{\min} < \frac{\partial M}{\partial P} < \left(\frac{\partial M}{\partial P}\right)_{\max}$$

If $\frac{\partial M}{\partial P}$ is outside the valid range, then the model should be modified, and the validation test is rerun until $\frac{\partial M}{\partial P}$ falls within the desired range.

Note that for first validation test, the parameter P can vary discretely, as well as continuously. However, for second validation test, the parameter P can only vary continuously because of the need to take a derivative with respect to P. Examples of continuously variable parameters for molecular modeling include temperature, pressure, and variables in a particular force field or solvent model. Examples of discretely variable parameters in a model include which atoms of the molecule are best modeled as rigid body

subunits, which complete solvent or force field model should be used for the molecular model system, and the presence/absence of other molecules, such as chaperons in the case of protein folding simulations.

Examples of measures *M* that may be used to test the validity of the molecular model include the potential energy of the molecule, the reaction forces and moments on the rigid bodies used to model collections of atoms, and the RMS (root mean square) deviation error in the static structures of a folded protein.

As stated previously, the present invention can be used to full advantage with modeling and simulation techniques which have a significant speed-up in the calculation of results, such as disclosed in the above-referenced co-pending U.S. applications. This should be evident from the description of the exemplary self-validation tests below.

Examples of Self-Validation Tests

A molecular model useful in simulations is one with multiple rigid bodies for different groups of constituent atoms of the subject molecule. The previously referenced co-pending patent applications describe an Order (N) torsion angle, rigid multibody systems which can simulate complex molecules. One assumption in this particular model is that the covalent peptide bonds and covalent bonds to the residue side chains of the subject molecule do not stretch or bend to any sufficient amount that would invalidate the motion behavior or shape of the molecule. To validate that these bonds do not stretch significantly, the internal reaction forces and torques at the bond locations during the entire dynamic simulation or at static solutions can be computed. If any of these reactions exceed the level necessary to stretch bonds beyond an acceptably small amount, then the particular molecular modeling solution may be invalid.

Tests to Validate Rigid Body Models

Fig. 2 illustrates the structure of a protein fragment with two residues, alanine dipeptide 150, in a rigid body model as described above. Alanine dipeptide has the amino acid formula of Ala-Ala, and the chemical formula of

$$NH_3^+ - CH - C_\alpha H - CH_3 - CONH - C_\alpha H - CH_3 - COO^-$$

where C_α are the alpha carbons in each residue and $CONH$ is the rigid peptide body 154 between each residue. The multibody system description contains seven bodies 151-157 with several atoms per body. Each body consists of one or more atoms that are considered rigidly attached together. The seven bodies represent a total of 23 atoms and the connections between the rigid bodies are covalent bonds represented as pin joints that allow the bodies to

rotate with respect to each other, but not to stretch or bend in other directions. Two of the pin joints on either side of the peptide body 151 are the configuration angles ϕ 156 and ψ 158.

Fig. 3A illustrates a reaction moment $\underline{M}$ 160 and reaction force $\underline{F}$ 162 and acting at the pin joint for the angle ϕ of the peptide body 151. In reality, the peptide body 151 may twist at an angle θ_{C-N} between the carbon atom (C) 161 and the nitrogen atom (N) 163, or stretch by a displacement r_{N-C_α} between the nitrogen atom (N) 163 and the alpha carbon (C_α) 165 if these reactions are too large. Self-validation test of the first type may be used since the collection of atoms assembled in each rigid body is discrete and not continuous.

Fig. 3B illustrates the method for refining the rigid body model if the reaction moment $\underline{M}$ exceeds the maximum allowed for the model. Here the discrete modeling parameter P is whether the peptide 151 should be considered as twisting into two rigid bodies or not, given the measure of the reaction moment M. In this test, the peptide body 151 is broken into two smaller bodies 151A and 151B with a new pin joint connecting the two bodies at an angle θ_{C-N} 166. The reaction moment $\underline{M}$ is projected onto the axis aligned with the pin joint for the angle θ_{C-N} and is the projected moment $\underline{M}_P$ 168. If the magnitude of the projected moment exceeds the magnitude of maximum allowable moment $\|\underline{M}_P\| > \|\underline{M}_{\max}\|$, then the pin joint for the angle θ_{C-N} 166 is added to the model along with the appropriate restoring moment, and the simulation is rerun.

Fig. 3C illustrates the method for refining the rigid body model if the reaction force $\underline{F}$ exceeds the maximum allowed for the model. In this example, the discrete modeling parameter is whether the peptide body 151 and a neighboring body 152 are displaced or not, given the measure M of the reaction force. For this test, a new slider joint connecting the two bodies 151 and 152 is created with the displacement r_{N-C_α} 172. The reaction force $\underline{F}$ 162 is projected onto the axis aligned with the slider joint for the displacement r_{N-C_α} and is the projected force $\underline{F}_P$ 170. If the magnitude of the projected force exceeds the magnitude of maximum allowable force $\|\underline{F}_P\| > \|\underline{F}_{\max}\|$, then the slider joint for the displacement r_{N-C_α} is added to the model along with the appropriate restoring force, and the simulation is rerun.

Fig. 4 is an exemplary graph of such a reaction force plotted for a simulation of alanine dipeptide with the model and integrators for the equations of the model's motions as described in the previously cited co-pending patent applications. If, after the molecular model settles at the final time, the reaction exceeds an allowable maximum time, then the model can be refined and rerun. This is another example of the present invention's self-validation method.

Tests to Validate Force Field and Solvent Models

The simulation of molecules and their behavior includes various approximations of the forces (including solvent forces) that actually move the molecules in nature. Extremely complex quantum mechanical formulations of inter-atomic forces are often approximated as simplified mathematical functions or truncated series expansions with parameters chosen by experimental observations. There are various force models available, such as "Amber," "Charmm," and "MM3" which are all different approximations of the same forces in nature. Since these forces cannot be known precisely, it is crucial to determine whether particular computational results are excessively dependent on unrealistic precision with respect to these models.

The parameters in these force models can be varied and rerun all or a portion of the molecular simulation to test internal consistency. A particular scalar function, or set of functions, which measures the deviation between two different solutions is specified, such as the RMS deviation between atom positions in two computed protein structures. When two solutions are provided by varying only a single parameter, the numerical value of this function gives the partial derivative, or the sensitivity of the solution with respect to that parameter. This so-called "sensitivity analysis" allows the determination of how robust or sensitive the folding path, final structure, and final potential energy is to changes in the force models or other parameters in the force models. In turn, this knowledge can be used to isolate particular parameters for refinement or to determine that a particular computation is unreliable.

These sensitivity analyses also apply to any other parameterized models used in the simulation of a molecule, such as the solvent model, temperature, and pH. With a measure function or set of measure functions, repeated runs of the simulation evaluate the sensitivity of the measure functions with respect to *continuously* changing individual modeling parameters. For example, sensitivity analyses can be applied to molecular dynamics and statics simulations of protein folding and ligand docking.

Fig. 5 illustrates such a *continuous* sensitivity analysis. The model is the alanine dipeptide protein fragment 150 of Fig. 2. The parameter varied is the initial value of the ψ angle 158 at the start of the simulation. The measure plotted is the magnitude of the final potential energy of the molecule after a static analysis is run to find a resting potential energy state. In this sensitivity analysis, the parameter is varied from -180° to $+180^{\circ}$. In this self-validation test, a particular force field and solvent model are tested to verify that the final potential energy of the molecule should stay the same regardless of the initial starting position of the atoms of the molecule. Since the initial starting angle can vary continuously, both the first and second types of self-validation tests can be performed. Note that the magnitude of the potential energy stays approximately the same, except near the value of $\psi = 0$. Thus a first type of validation test shows that the model breaks down by plotting the measure of the potential energy, E, for all starting values of ψ. A second type of validation test shows that the change of E with respect to ψ near $\psi = 0$ is too large, i.e., $\frac{\partial E}{\partial \psi} > \left(\frac{\partial E}{\partial \psi}\right)_{\max}$. The information gleaned from this plot is used to refine the force field and solvent model.

Test to Validate Structural Insensitivity

The present invention can also be used to test the molecular model for the structural insensitivity. For example, a folded protein structure or its response to a ligand docking is often insensitive to the actual sequence of residues in certain regions of the amino acid sequence. Thus, slight genetic variations in the amino acid sequence do not change the form or function of the protein. In accordance with the present invention, the simulation of the motions of the molecular model allows the variation of the residues to determine the sensitivity of the folding path, final structure, ligand docking, and potential energy to changes in residue sequence. In this embodiment of the present invention, the parameter being changed is *discrete*, such as the amino acid sequence rather than a continuous force field or other modeling parameter.

Fig. 6 illustrates such a *discrete* sensitivity analysis. The model is related to the alanine dipeptide protein fragment 150 of Fig. 2. The parameter varied is the second residue in the dipeptide. That is, the second alanine is replaced with an Arginine, then an Aspartine, etc. The measure plotted is the final potential energy of the molecule after a static analysis is run to find a minimum potential energy state. In this sensitivity analysis, the parameter is *discretely* varied from one amino acid residue to another amino acid residue. In

this exemplary self-validation case, the magnitude of the final potential energy E for each case is within the tolerance for the model used: $E_{min} < E < E_{max}$.

Conclusion

Of course, one way of testing a molecular model is to compare the simulation solutions to known native protein structures or drug ligand binding (as experimentally determined by X-ray crystallography or NMR (nuclear magnetic resonance) techniques). The present invention also allows for the validation of protein structures and ligand bindings determined by simulation, but not yet experimentally analyzed. The present invention exploits features of high-speed simulation methods to test the validity of certain approximations employed in the modeling process, namely the rigid body assumption, the force and solvent models, and of proteins with the particular amino acid sequence specified. In accordance with other aspects of the present invention, the sensitivity of molecular models to changes in a particular amino acid sequence can be applied to the field of protein design. For example, by modeling proteins with insensitivities to certain parameters, such as temperature or pH, actual proteins can be modified for stability in certain applications, such as detergent enzymes.

As a further example of self-validation, the present methods allow one to assess the significance of estimates of partial charge present on atoms of a molecular system. For example, the partial charge on an atom might be estimated as 0.5 electron units +/- 0.1. The molecular system is simulated based on this selected parameter, and a value of a validation measure, such as a binding affinity, is determined from the resulting model. The simulation is then repeated using a different value for the charge on the atom within the margin of error (e.g., 0.6 electron units) and a second value of the validation measure is determined. If the validation measure does not change significantly, then one has an indication that the model is reliable, as is the binding affinity calculated. If, however, the binding affinity changes significantly (e.g., by a factor of 2), then one knows that the binding affinity determined may be significantly in error, and that the model needs refinement.

As a further example, the model parameter can be the identity of an atom or group of atoms within a molecule of the molecular system. For, example, the model parameter can be the identity of an amino acid. The molecular system can be simulated for different amino acids. If the model is accurate, one would expect that conservative changes between amino acids (see e.g., Stryer, Biochemistry(Freeman , NY, 4th Ed 1995), would result in relatively minor changes of a validation measure, and that nonconservative changes

would result in larger changes. In such measures, the validation measure can be a quantitative value, such as the binding affinity of the protein for a target, or a qualitative result such as the shape of the folded protein. Although testing of a single position is not necessarily conclusive, in general, if conservative substitutions give rise to smaller changes than nonconservative substitutions in the validation measure, then one has reason to think the computer modeling of the molecular simulation is accurate.

As a further example, the model parameter can be the temperature of the molecular system. The molecular system can be simulated for different temperatures, and validations measurements made at the different temperatures. In this case, a suitable validation measure is the velocity of atoms in the molecular system. If the model is accurate, the velocity of atoms should increase, as does the temperature.Similarly, if the model parameter is pressure, the velocity of atoms should also increase with increasing pressure.

If a model survives the test of self-validation, then validation measures of the model can also be compared with experimentally determined measures of the same system as a further check. However, by first performing a self-validation, the need for chemical synthesis and chemical or biochemical assays to perform validation is at least reduced.

The validated molecular modeling system can then be used in various applications including screening libraries of compounds for interaction with a target, as discussed in Background section. Compounds that appear to have the desired interaction with the target identified by molecular modeling can then be synthesized chemically and tested in biochemical assays.

Therefore, while the foregoing is a complete description of the embodiments of the invention, it should be evident that various modifications, alternatives and equivalents may be made and used. Accordingly, the above description should not be taken as limiting the scope of the invention which is defined by the metes and bounds of the appended claims.